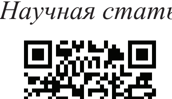
ИНФОРМАТИКА, ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА И УПРАВЛЕНИЕ
INFORMATION TECHNOLOGY, COMPUTER SCIENCE, AND MANAGEMENT





УДК 681.3
https://doi.org/10.23947/2687-1653-2022-22-1-57-66

Научная статья



Проблема обеспечения производительности доверенных систем управления с глубинным обучением

А. А. Зеленский ✉, **Т. Х. Абдулин**, **М. М. Жданова**, **В. В. Воронин**, **А. А. Грибков**
Московский государственный технологический университет «СТАНКИН» (Москва, Российская Федерация)
✉ Zelenskyaa@gmail.com

Введение. Рассмотрены значение машинного обучения в условиях цифровой трансформации промышленности, методы реализации глубинного обучения для обеспечения производительности доверительных систем управления. Определена необходимость использования для глубинного машинного обучения сверточных искусственных нейронных сетей. Кратко рассмотрены различные технологии и архитектуры реализации искусственных нейронных сетей, проведен сравнительный анализ их производительности. Целью работы является исследование необходимости разработки новых подходов к архитектуре вычислительных машин для решения задач глубинного машинного обучения при реализации доверительной системы управления.
Материалы и методы. В условиях цифровой трансформации использование искусственного интеллекта выходит на новый уровень. В основе технической реализации искусственных нейронных систем с глубинным машинным обучением лежит использование одной из трех базовых технологий: высокопроизводительных вычислений (HPC) с параллельной обработкой данных, нейроморфных вычислений (NC) и квантовых вычислений (QC).
Результаты исследования. Проведен анализ моделей реализации глубинного машинного обучения, базовых технологий и архитектуры вычислительных машин, а также требований по обеспечению доверия к системам управления, использующих глубинное машинное обучение. Выявлена проблема дефицита вычислительных мощностей для решения таких задач. Ни одна из существующих технологий не позволяет решать полный комплекс задач обучения и импеданса. Современный уровень технологий не обеспечивает информационной безопасности и надежности работы нейронных сетей. Практическая реализация доверенных систем управления с глубинным машинным обучением на базе имеющихся технологий для существенной части задач не обеспечивает достаточной производительности.
Обсуждение и заключения. Проведенное исследование позволило выявить проблему дефицита вычислительных мощностей для решения задач глубинного машинного обучения. На основе анализа требований к доверенным системам управления определены объективные сложности их реализации на базе существующих технологий и установлена необходимость разработки новых подходов к архитектуре вычислительных машин.

Ключевые слова: глубинное машинное обучение, процессор, доверенная система, информационная безопасность, вычислительная машина, искусственный интеллект.

Для цитирования: Проблема обеспечения производительности доверенных систем управления с глубинным обучением / А. А. Зеленский, Т. Х. Абдулин, М. М. Жданова [и др.] // Advanced Engineering Research. — 2022. — Т. 22, № 1. — С. 57–66. https://doi.org/10.23947/2687-1653-2022-22-1-57-66

Финансирование: Работа выполнена по гранту Российского научного фонда No 21-79-10392, https://rscf.ru/project/21-79-10392/

Original article

Challenge of the performance management of trust control systems with deep learning

Aleksandr A. Zelensky ✉, **Tagir K. Abdullin**, **Marina M. Zhdanova**, **Vyacheslav V. Voronin**, **Andrey A. Gribkov**
Moscow State Technological University "STANKIN" (Moscow, Russian Federation)
✉ Zelenskyaa@gmail.com

Introduction. The significance of machine learning under the conditions of digital transformation of industry, and methods of implementing deep learning to provide the performance of trust management systems are considered. The necessity of using convolutional artificial neural networks for deep machine learning is determined. Various technologies and architectures for the implementation of artificial neural networks are briefly considered; a comparative analysis of their performance is carried out. The work objective is to study the need to develop new approaches to the architecture of computing machines for solving problems of deep machine learning in the trust management system implementation.

Materials and Methods. In the context of digital transformation, the use of artificial intelligence reaches a new level. The technical implementation of artificial neural systems with deep machine learning is based on the use of one of three basic technologies: high performance computing (HPC) with parallel data processing, neuromorphic computing (NC), and quantum computing (QC).

Results. Implementation models for deep machine learning, basic technologies and architecture of computing machines, as well as requirements for trust assurance in control systems using deep machine learning are analyzed. The problem of shortage of computation power for solving such problems is identified. None of the currently existing technologies can solve the full range of learning and impedance problems. The current level of technology does not provide information security and reliability of neural networks. The practical implementation of trust management systems with deep machine learning based on existing technologies for a significant part of the tasks does not provide a sufficient level of performance.

Discussion and Conclusions. The study made it possible to identify the challenge of the computation power shortage for solving problems of deep machine learning. Through the analysis of the requirements for trust management systems, the external challenges of their implementation on the basis of existing technologies, and the need to develop new approaches to the computer architecture are determined.

Keywords: deep machine learning, processor, trust system, information security, computer, artificial intelligence.

For citation: A. A. Zelensky, T. K. Abdullin, M. M. Zhdanova, V. V. Voronin, A. A. Gribkov. Challenge of the performance management of trust control systems with deep learning. Advanced Engineering Research, 2022, vol. 22, no. 1, pp. 57–66. (In Russ). https://doi.org/10.23947/2687-1653-2022-22-1-57-66

Funding information: the research is done on the Russian Science Foundation grant no. 21-79-10392, https://rscf.ru/project/21-79-10392/

Введение. В последние 9 лет в мире, и в России в т. ч., все более масштабно совершается 4-я промышленная революция, одной из ключевых составляющих которой является цифровая трансформация, затрагивающая все аспекты экономической жизни — от крупного промышленного производства до сферы обслуживания, науки, образования и домашних хозяйств. В этих условиях применение машины в задачах, с которыми человек не справляется или справляется хуже машины, существенно расширяется. Если раньше речь шла о механизации ручного труда и автоматизации производства, то теперь человек может быть заменен машиной при решении задач обработки данных, анализа, прогнозирования и управления различными системами: оборудованием, технологическими процессами, промышленными предприятиями, торговыми сетями и др.

Практической основой замены человека машиной в отдельных областях интеллектуальной деятельности является использование искусственного интеллекта. Создание «сильного» искусственного интеллекта — задача будущего, связанная с необходимостью создания и развития новых технологий, а также решением значимых этических проблем. В настоящее время доступны для использования системы машинного обучения различной сложности, представляющие собой ступень к созданию «сильного» искусственного интеллекта.

Мировой рынок систем машинного обучения активно расширяется. В 2020 г. его объем составил 11,33 млрд $, в 2021 г. он вырос до 15,50 млрд $, а к 2028 г. достигнет уровня 152,24 млрд $, показывая среднегодовой рост на 38,6 %[1].

Сфера применения систем машинного обучения очень велика и включает в себя маркетинг и торговлю, банковскую деятельность, промышленное производство, медицину и др. Системы машинного обучения наиболее востребованы в следующих областях промышленности:

— в робототехнике для интеллектуализации промышленных и сервисных роботов, в т. ч. коллаборативных;

— в автоматизированных системах управления технологическими процессами и предприятиями;

— в системах управления производственными процессами;

— в системах управления поставками и взаимоотношениями с клиентами;

— в исполнительных системах производства;

— в системах производственной аналитики технологического оборудования;

— в системах бизнес-аналитики и др.

Сложные технологические системы, например реализуемые в конструкциях металлорежущих станков с числовым программным управлением в режиме реального времени, в настоящее время не могут быть оснащены системами управления, пригодными для машинного обучения. Для этого требуются вычислительные мощности, способные выполнять существенные по объему вычисления в течении десятков микросекунд. Такая задача с помощью современных технических средств не может быть решена. Поэтому в большинстве случаев процесс машинного обучения системы управления осуществляется на отделенных от нее вычислительных машинах без ограничений времени, а затем результаты обучения передаются в систему управления в виде рекомендаций, указаний по режимам работы, смене инструмента, интервалам поверки и т. д.

Методы машинного обучения условно соответствуют видам умозаключений, лежащих в их основе: индукции, дедукции и традукции. Метод контролируемого обучения по прецедентам, когда в машину загружаются большие объемы данных, предварительно маркированных оператором-человеком, соответствует индукции. Метод обучения без учителя, когда машина должна сама найти закономерности в данных, выявить шаблоны, упорядочить и структурировать данные, соответствует индукции и традукции. Дедукции и традукции соответствует экспертный метод, основанный на использовании для обработки данных заданных закономерностей и шаблонов. Традукция реализуется в основном через использование трансферного обучения, основанного на применении к данной задаче знаний, полученных при решении другой задачи.

Для машинного обучения используются различные технологии и математические модели. Наибольший потенциал развития имеет модель искусственных нейронных сетей (ИНС), построенная по аналогии с биологическими нейронными сетями, т. е. сетями нервных клеток живого организма. ИНС представляет собой систему соединённых между собой и взаимодействующих искусственных нейронов, реализуемых в виде процессоров, процессорных элементов в виде ускорителей или сопроцессоров под управлением центрального процессора. Нейроны ИНС располагаются по уровням (слоям). Первый уровень соответствует получению, обработке входных данных и передаче их на следующий уровень. Промежуточные уровни — скрытые, их задача обрабатывать поступающие данные и передавать на последний (выходной) уровень. В нейронной сети может быть несколько скрытых уровней, перемежающихся с уровнями, на которых выполняются логические, математические и другие преобразования. От уровня к уровню данные обрабатываются, на каждом последующем уровне выявляются взаимосвязи предыдущего. Такая многоуровневая ИНС имеет большие возможности и может быть использована для реализации глубинного машинного обучения [1–3].

Глубинное машинное обучение — метод проектирования ИНС с помощью многослойных фильтров для извлечения и моделирования признаков из набора входных данных[2]. Такое обучение может быть контролируемым и неконтролируемым. Также возможно использование глубинного машинного обучения для экспертных систем.

В основе технической реализации искусственных нейронных систем с глубинным машинным обучением лежит использование одной из трех базовых технологий: высокопроизводительных вычислений с параллельной обработкой данных, нейроморфных и квантовых вычислений[3].

[1] Machine Learning Market, 2021-2028 // Hardware & Software IT Services Market Research Report. — 2021. — P. 160. — URL: https://www.fortunebusinessinsights.com/infographics/machine-learning-market-102226 (дата обращения: 06.11.2021).

[2] Глек, П. Глубокое обучение (Deep Learning): краткий туториал // neurohive.io : [сайт]. URL: https://neurohive.io/ru/osnovy-datascience/glubokoe-obuchenie-deep-learning-kratkij-tutorial/ (дата обращения: 06.11.2021).

[3] Как сократить издержки при использовании ИИ / HitachiVantaraCorporation : [сайт]. URL: https://hitachi.cnews.ru/articles/2021-06-14_kak_sokratit_izderzhki_pri_ispolzovanii (дата обращения: 07.11.2021).

Материалы и методы

Высокопроизводительные вычисления. Высокопроизводительные вычисления с параллельной обработкой данных реализуются посредством гибридных вычислительных систем, т. е. систем с гетерогенной аппаратной вычислительной структурой, включающей в себя центральный процессор (CPU) и дополнительный вычислительный модуль в виде ускорителя или сопроцессора. В зависимости от процессоров, используемых для параллельной обработки данных, гибридные вычислительные машины имеют одну из четырех архитектур:

1. Архитектура на базе графических процессоров (GPU). Наиболее распространенные решения представляют собой графические ускорители, расширяющие вычислительные возможности центрального процессора компьютерной системы. Последние достижения в данной области — графические ускорители NVIDIA Tesla V100, обеспечивающие для задач глубинного машинного обучения производительность 120 TFLOPS, т. е. 1,2×1014 операций с плавающей запятой в секунду[4]. Это в 500–1000 раз выше производительности обычного персонального компьютера (ПК). Необходимо также учитывать, что указанная производительность обеспечивается при решении задач, требующих значительных вычислительных мощностей и существенных затрат времени, но не при работе в режиме реального времени. Архитектура на базе графических процессоров в настоящее время наиболее доступна. В частности, для реализации системы с ограниченной вычислительной мощностью достаточно иметь на ПК видеокарту с графическим процессором NVidia, реализующую программно-аппаратную архитектуру параллельных вычислений CUDA. Наряду с CUDA к числу технологий GPGPU, использующих графический процессор видеокарты для компьютерной графики в целях производства математических вычислений, относится технология AMD FireStream (для графических процессоров ускорителей ATI). Мировой рынок графических процессоров в настоящее время составляет около 26 млрд $, темп роста рынка — до 34 % в год[5].

2. Архитектура на базе программируемых пользователем вентильных матриц (FPGA) — полупроводниковых устройств, которые могут перепрограммироваться и менять топологию соединений в процессе использования. Номинальная производительность данных устройств сравнительно невысока — около 20 TFLOPS, однако эффективность использования вычислительных мощностей наибольшая среди всех рассмотренных архитектур. Она в 6–7 раз выше, чем у графических ускорителей. Высокая эффективность FPGA обусловлена гибкостью и скоростью подстройки под решаемые вычислительные задачи. Мировой рынок FPGA по данным компании Grand View Research, в 2020 г. составил 9,85 млрд $, ожидаемые темпы роста рынка на период до 2027 г. — 9,7 % в год[6].

3. Архитектура на базе интегральных схем специального назначения (ASIC). Благодаря узкой специализации решаемых вычислительных задач они могут быть существенно проще, дешевле и компактнее. Производительность ASIC может достигать 1000 TFLOPS, однако эффективность использования вычислительных мощностей, например количество распознанных изображений, примерно в 2 раза ниже, чем в графических ускорителях. Темпы роста рынка ASIC существенно ниже, чем рынка графических процессоров. В 2020 г. по данным компании Global Industry Analysts мировой рынок ASIC составил 17,3 млрд $, ожидаемый среднегодовой рост рынка до 2027 г. — 7,7 %[7].

4. Архитектура на базе однокристальных ускорителей (SoC). «Система на кристалле» SoC — полнофункциональное электронное устройство, имеющее материнскую плату, процессор и другие необходимые для работы компоненты, размещённое на одной интегральной схеме. SoC распространены в мобильных компьютерах (смартфонах), одноплатных компьютерах и других встроенных системах. При этом SoC имеют существенный потенциал использования в составе гибридных вычислительных машин. Кроме того, возможны решения в виде однокристальной сборки SoC с элементами FPGA (архитектура Versal для адаптивных вычислений компании Xilinx[8]). Рынок SoC в настоящее время очень велик и за 2020 г. составляет 79,7 млрд $. На период до 2027 г. прогнозируется рост рынка на 4,4 % в год, к 2027 г. объем рынка достигнет 107,4 млрд $[9].

На рис. 1, 2 приведены данные по фактическому и прогнозируемому росту мирового рынка чипов для глубинного машинного обучения, подготовленные консалтинговой компанией Omdia[10].

---

[4] Как сократить издержки при использовании ИИ Указ. соч.

[5] Global Graphics Processing Unit (GPU) Market Insights and Forecast to 2027 // QYResearch. — 2021. — P. 116. — URL: https://reports.valuates.com/market-reports/QYRE-Auto-25V3358/global-graphics-processing-unit-gpu (дата обращения: 10.11.2021).

[6] Field Programmable Gate Array Market, 2020 — 2027 // Grand View Research. — 2020. — P. 130. — URL: https://www.marketresearch.com/Grand-View-Research-v4060/Field-Programmable-Gate-Array-Size-13223073/ (дата обращения: 11.11.2021).

[7] ASIC — Global Market Trajectory & Analytics April 2021// Global Industry Analysts, Inc. — 2021. — P. 185. — URL: https://www.researchandmarkets.com/reports/5140939/asic-global-market-trajectory-and-analytics дата обращения: 11.11.2021).

[8] Xilinx Unveils Its Most Ambitious Accelerator Platform // Michael Feldman. — 2018. — URL: https://www.top500.org/news/xilinx-unveils-its-most-ambitious-accelerator-platform/ (дата обращения: 10.11.2021).

[9] System-On-A-Chip (SoC) — Global Market Trajectory & Analytics. Report April 2021 // Global Global Industry Analysts, Inc. – 2021. — URL: https://www.researchandmarkets.com/reports/2832316/system-on-a-chip-soc-global-market-trajectory (дата обращения: 11.11.2021).

[10] Joshi, A. Deep Learning Chipsets Report — 2020 // Omdia Market. — 2020. — URL: https://omdia.tech.informa.com/products/deep-learning-chipsets-report---2020

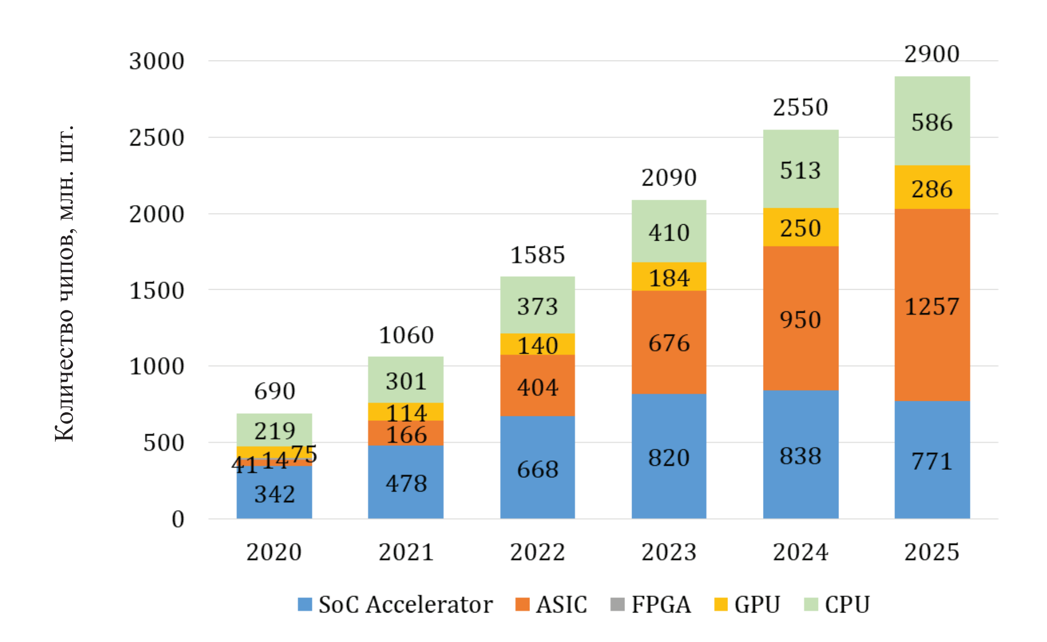


Рис. 1. Динамика роста количества чипов для глубинного обучения по годам

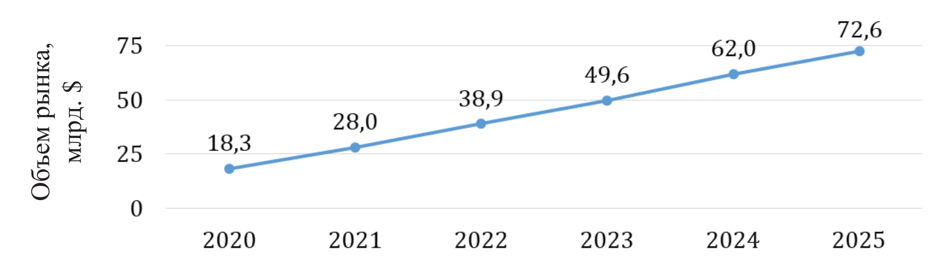


Рис. 2. Динамика изменения рынка чипов для глубинного обучения по годам[11]

Анализ показывает, что на фоне общего роста рынка наибольшие перспективы имеют ASIC, значимые позиции сохранят также GPU и SoC. Ускорители GPU на достаточно длительную перспективу не имеют адекватной замены в решении сложных задач, в т. ч. в процессе обучения, а SoC незаменимы для мобильных реализаций систем глубинного машинного обучения, а также для параллельных вычислений с целью разгрузки центрального процессора. Отказ от FPGA в системах глубинного машинного обучения представляется необязательным, хотя доля таких процессоров, вероятно, будет сравнительно небольшой.

Одним из наиболее перспективных направлений развития гибридных вычислительных машин является использование тензорных и других специализированных сопроцессоров, например процессоров машинного зрения. Такие сопроцессоры могут строиться на базе наиболее распространенных и производительных ASIC, а также FPGA или GPU. Отличие сопроцессоров от ускорителей заключается в степени интеграции с центральным процессором. Центральный процессор посредством специальной области памяти транслирует управляющие инструкции на ускоритель. Сопроцессор отслеживает поток инструкций машинного кода, поступающий из оперативной памяти в центральный процессор, и перехватывает инструкции, соответствующие его функциональному назначению, например задачи тензорных преобразований, распознавания образов и т. д. Для решения масштабных задач, требующих длительных распределенных вычислений, целесообразно применение ускорителя, для частого и многократного выполнения несложных вычислительных задач — применение сопроцессора, при котором центральный процессор не загружается и не замедляет обработку данных.

Эксплуатационные свойства вычислительных машин существенно зависят от используемой архитектуры. Для решения задач глубинного машинного обучения, куда затрачиваются до 80 % вычислительных мощностей, наилучшим образом подходят машины на базе графических ускорителей. Они имеют высокую производительность при решении сложных задач, требующих существенных затрат времени, обладают высокой гибкостью и максимальной точностью вычислений, но низкой относительной производительностью. Например, для процессоров NVidia она составляет 1,3–1,8 GOPS/W. Вычислительные

[11] Joshi, A. Deep Learning Chipsets Report — 2020 // Omdia Market. — 2020. — URL: https://omdia.tech.informa.com/products/deep-learning-chipsets-report---2020 (accessed: 12.11.2021).

машины на базе ASIC обладают максимальной абсолютной и относительной производительностью. Например, для процессоров neuIBM относительная производительность составляет 254 GOPS/W [4]. Однако такие машины имеют низкую гибкость и ограниченную точность, поэтому их целесообразно использовать при решении типовых, например матричных или тензорных преобразований, повторяющихся или многопоточных задач в режиме реального времени.

Вычислительные машины на базе FPGA обладают высокими параметрами гибкости, точности, абсолютной и относительной производительности. Например, для процессоров Tegra TX1 относительная производительность составляет 70 GOPS/W Однако такие машины имеют сравнительно высокую стоимость, поэтому их целесообразно использовать для научных целей, когда требуется лишь несколько вычислительных машин заданной конфигурации, а также для разработки архитектуры массово выпускаемых процессоров ASIC и SoC.

Несмотря на совершенствование архитектуры вычислительных машин потенциал роста мощностей для высокопроизводительных вычислений в скором времени будет исчерпан. Число транзисторов на кристалле за последние 5 лет увеличилось примерно в 12 раз [5], а объем вычислений в процессе машинного обучения — в 150 тыс. раз[12].

Нейроморфные вычисления. Возможным способом устранения нехватки вычислительных мощностей для искусственных нейронных систем с глубинным машинным обучением является использование нейроморфных вычислений и соответствующих чипов. Нейроморфный чип — это процессор, работа которого основана на принципах человеческого мозга. Такой чип моделируют работу нейронов и их отростков — аксонов и дендритов, отвечающих за передачу и восприятие данных. Связи между нейронами образуются за счет синапсов — специальных контактов, по которым транслируются электрические сигналы.

К числу наиболее известных разработок в данной области относятся нейроморфные процессоры TrueNorth компании IBM и процессоры Loihi компании Intel. В них используется асинхронная кластерная архитектура и модель сверточной нейронной сети — однонаправленная многослойная сеть с чередованием свёрточных и субдискретизирующих слоёв. Процессор TrueNorth выполнен на базе технологий 28 нм, Loihi — на базе 14 нм [6].

Многопроцессорная система TrueNorth NS16e-4, содержащая 100 млн нейронов и предназначенная для работы с сетями для глубинного машинного обучения, была представлена компанией IBM в 2018 г. [7]. Каждый чип содержит 1 млн цифровых нейронов и 256 млн синапсов, заключенных в 4096 синапсных ядрах; энергопотребление каждого чипа — 70 мВт.

Процессор Loihi, представленный в 2017 г., содержит 131 тыс. искусственных нейронов и 131 млн синапсов. В 2019–2020 гг. Intel представила два продукта на основе Loihi — PohoikiBeach и PohoikiSprings. Вычислительная система PohoikiBeach, включающая 64 процессора Loihi, суммарно располагает 8,32 млн нейронов и 8,32 млрд синапсов. Вычислительная система PohoikiSprings включает 768 процессоров Loihi, имеет 100 млн нейронов и 100 млрд синапсов[13].

В России работы по созданию нейроморфных процессоров ведутся уже несколько лет. В 2020 г. компания «Мотив Нейроморфные Технологии» создала нейрочип «Алтай» [8]. Технологическая норма процессора — 28 нм, энергопотребление — около 0,5 Вт, площадь кристалла — 64 мм^2 (для сравнения: TrueNorth — 430 мм^2, Loihi — 60 мм^2). В нем 131 тыс. нейронов, между ними — 67 млн связей.

Для оценки качества нейроморфных процессоров используют:

1. Абсолютный показатель производительности. Это количество миллиардов выполняемых синоптических операций в секунду (GSOPS).

2. Показатель энергоэффективности. Это количество пикоджоулей энергии, затраченной на выполнение одной осинаптической операции (pJ/SOP).

Процессор TrueNorth имеет производительность 58 GSOPS и энергоэффективность 26 pJ/SOP[14]. Аналогичную энергоэффективность (23,7 pJ/SOP [9]) имеет процессор Loihi.

Единственным конкурентом нефроморфных процессоров в реализации нейронных сетей с глубинным машинным обучением на среднесрочную перспективу (8–12 лет) являются гибридные вычислительные машины с сопроцессорами ASIC. Такие процессоры имеют меньшую, но сопоставимую производительность выполнения синаптических операций и более высокую энергоэффективность. В частности, процессор ASIC,

[12] Thompson, N.C., Greenewald, K., Lee, K., Manso, G.F. The Computational Limits of Deep Learning// arXiv preprint arXiv:2007.05558. – 2020. (дата обращения: 12.11.2021)
[13] Intel Scales Neuromorphic Research System to 100 Million Neurons// Intel. — 2020. — URL: https://newsroom.intel.com/news/intel-scales-neuromorphic-research-system-100-million-neurons/ (дата обращения: 12.11.2021).
[14] Нейроморфный процессор «Алтай» // motivnt.ru : [сайт]. URL: https://motivnt.ru/neurochip-altai/ (дата обращения: 10.11.2021).

описанный в работе [10], обеспечивает производительность выполнения синаптических операций 8,7 GSOPS и энергоэффективность 15,2 pJ/SOP.

Мировой рынок нейроморфных чипов сформировался сравнительно недавно и поэтому невелик. В 2020 г. его объем составил всего 22,5 млн $. При этом темпы роста рынка очень высоки. К 2026 г. рынок вырастет до 333,6 млн $, что соответствует среднегодовому росту на 47,4 %[15].

Иногда под нейроморфными чипами понимают все виды процессоров, внешне воспроизводящих работу нейронов независимо от внутренней структуры технического устройства, которое может не соответствовать характеру взаимодействия нейронов. Такие процессоры, используемые для построения искусственных нейронных сетей, правильно называть нейронными. К нейронным процессорам, наряду с нейроморфными чипами, также относятся процессоры (чипы) с тензорными и другими специализированными сопроцессорами машинного зрения, распознавания речи и др. Мировой рынок нейронных процессоров в настоящее время составляет 2,3 млрд $, а к 2027 г. вырастет до 10,4 млрд $, т. о. средний рост составит 24,2 % в год[16].

Квантовые вычисления. В долгосрочной перспективе средством устранения дефицита вычислительных мощностей может стать развитие квантовых вычислений. Квантовые вычисления — решение задач с помощью манипуляции квантовыми объектами: атомами, молекулами, фотонами, электронами и специально созданными макроструктурами. Манипуляции квантовыми объектами дают возможность использования:

— квантовой суперпозиции, которая проявляется в способности квантовых систем одновременно находиться во всех возможных состояниях;

— квантовой запутанности, которая проявляется в сильной взаимосвязи между параметрами специальным образом приготовленных квантовых систем.

Устройства для квантовых вычислений принято разделять на два больших класса [11]: универсальные квантовые компьютеры и квантовые симуляторы. Первые подобно центральным процессорам могут решать любую алгоритмическую задачу, квантовые симуляторы — это аналоговые компьютеры для решения узкоспециализированных задач.

Технологии создания универсальных квантовых компьютеров в настоящее время находятся на стадии формирования. Создаваемые вычислительные машины демонстрируют «квантовое превосходство» в решении отдельных задач, но пока не могут быть использованы для формирования искусственных нейронных сетей с глубинным машинным обучением. Наибольшую активность в создании квантового компьютера демонстрируют компании:

1. Google. В 2018 г. построен 72-кубитный квантовый процессор Bristlecone, в 2019 г. — более точный 53-кубитный квантовый процессор Sycamore.
2. Intel. В 2018 г. построена 49-кубитная сверхпроводящая квантовая микросхема TangleLake.
3. IBM. В 2017 г. создан и протестирован 50-кубитный квантовый процессор, в 2019 г. — первый в мире коммерческий 20-кубитный квантовый компьютер IBM Q SystemOne и др.

Единственным адиабатическим квантовым компьютером, представленным на рынке, является D-WaveSystems, выпускаемый в вариантах от 16 до 2000 кубитов, организованных в кластеры по 8 кубитов в каждом.

Область квантовых симуляторов также активно развивается. Один из наиболее сложных симуляторов такого типа представлен совместной разработкой 2017 г. Мэрилендского университета и Национального института стандартов и технологий (США). Этот 53-кубитный симулятор использует в качестве кубитов холодные ионы иттербия. Аналогичный по возможностям 51-кубитный квантовый симулятор на базе атомов рубидия был разработан группой ученых Гарвардского университета и Массачусетского технологического института.

В России также реализуется ряд проектов, развивающих технологии квантовых вычислений. В частности, уже несколько лет ведётся разработка сверхпроводящего процессора учёными консорциума, куда входят МИСиС, ИФТТ РАН, МГТУ им. Н. Э. Баумана, ВНИИА им. Духова и другие организации. На сегодняшний день консорциумом отлажена технология изготовления сверхпроводниковых двухкубитных схем,

[15] Neuromorphic Chip Market — Growth, Trends, COVID-19 Impact, and Forecasts (2021–2026) // Mordor Intelligence. — 2020. — URL: https://www.mordorintelligence.com/industry-reports/neuromorphic-chip-market (дата обращения: 11.11.2021).

[16] Neuromorphic Chips — Global Market Trajectory & Analytics // Global Industry Analysts, Inc. 2021. 118 p. URL: https://www.researchandmarkets.com/reports/4805280/neuromorphic-chips-global-market-trajectory-and (дата обращения: 10.11.2021).

экспериментально охарактеризованы и продемонстрированы двухкубитные логические вентили, осуществляющие квантовое запутывание, необходимое для работы квантовых процессоров. Достоверность логических операций находится в пределах 85–95 %.

В 2020 г. в квантовые вычисления в мире было вложено 675 млн $, что более чем в 3 раза превышает объем инвестиций 2019 г. (211 млн $). В 2021 г. объем инвестиций в квантовые вычисления превышает 800 млн $ [12].

Доверительные системы управления. Одним из базовых требований к системам управления, в том числе к производственным, является обеспечение к ним необходимого уровня доверия. Согласно ГОСТ Р 54583-2011 «Методы и средства обеспечения безопасности. Основы доверия к безопасности информационных технологий. Часть 3 — Анализ методов доверия» целью обеспечения доверия является создание уверенности в надёжном функционировании продукта в заданных условиях. Для обеспечения указанного доверия информационная система должна обладать следующими эксплуатационными свойствами [13]:

— функциональной надёжностью, т. е. способностью выполнять свою функцию с заданной достоверностью, которая в свою очередь нормируется числом отказов, погрешностью и повторяемостью результатов вычислений;

— информационной безопасностью, т. е. способностью обеспечивать заданный уровень конфиденциальности, доступности и целостности информации: хранимой, передаваемой, принимаемой и обрабатываемой в процессе работы системы.

Предметом настоящей работы являются системы управления, которые относятся к информационным. Поэтому для них также справедливы указанные выше требования к эксплуатационным свойствам. Однако у систем управления имеется своя специфика в определении доверия. Доверенная система управления должна обладать:

— способностью управлять, например роботом, станком, предприятием и т. д. по заданному числу параметров с заданной надежностью, которая регламентируется числом отказов, погрешностью и повторяемостью, и с заданной производительностью, которая в свою очередь регламентируется временем обработки данных и выполнения управляющих команд;

— возможностью контроля элементов, структуры и процессов системы на аппаратном и программном уровне с целью обеспечения информационной безопасности.

Если система управления имеет функцию глубинного машинного обучения, то выполнение первого из указанных требований накладывает жесткие ограничения на используемые средства технической реализации. Это должна быть оптимальная для формирования сверточной нейронной сети вычислительная машина с высокими параметрами производительности, точности и надежность вычислений.

Если не рассматривать вариант использования квантовых процессоров, полнофункциональных реализаций которых пока нет, то первому требованию в полной мере не соответствуют ни гибридные вычислительные машины на базе всех рассмотренных архитектур, ни вычислительные машины на базе нейроморфных процессоров. Вычислительные машины на базе ASIC и нейроморфных процессоров не обеспечивают высокой точности и надежности, а гибридные вычислительные машины с ускорителями GPU или SoC неоптимальны для работы в режиме реального времени, в том числе импеданса.

Определенный компромисс обеспечивается при использовании гибридных вычислительных машин на базе FPGA, однако такие машины имеют высокую стоимость при серийном производстве, существенно меньшую производительность, чем машины с ASIC, и существенно меньшие возможности сложных вычислений, чем машины с GPU. Другим вариантом достижения компромисса является одновременное использование центрального процессора с графическим ускорителем для решения сложных задач в процессе машинного обучения и тензорных или иных узкоспециализированных сопроцессоров на базе ASIC для обработки данных в режиме реального времени.

Второе требование, хотя и является техническим по содержанию, на практике выступает в качестве экономического. Реализация системы управления технологическим оборудованием с глубинным машинным обучением возможна только посредством сверточной нейронной сети, контроль работы которой извне не представляется возможным. Информационная безопасность может быть обеспечена только при условии, что основная часть вычислительной машины будет создаваться отечественными производителями, которые прошли сертификацию в области информационной безопасности.

В настоящее время в России основная часть систем управления построена на основе зарубежных микроэлектронных компонентов. Доля таких компонентов превышает 85 % [14]. Обеспечение информационной безопасности в случае использования в вычислительных машинах импортных компонентов не имеет однозначного решения и зависит от структуры создаваемой искусственной нейронной сети и порядка ее

использования. В частности, при использовании гибридных вычислительных машин информационная безопасность может быть существенно повышена за счет локализации передачи данных между центральным процессором и ускорителем или сопроцессором.

Результаты исследования. Проведенный анализ моделей глубинного машинного обучения, базовых технологий и архитектуры вычислительных машин, а также требований по обеспечению доверия к системам управления, использующих глубинное машинное обучение, позволяет сделать следующие выводы:

1. Имеется объективная проблема дефицита вычислительных мощностей для решения задач глубинного машинного обучения. Ни одна из существующих в настоящее время технологий не позволяет решать полный комплекс задач обучения и импеданса.

2. Поскольку глубинное машинное обучение реализуется на базе модели сверточных нейронных сетей, их внешний контроль с целью обеспечения информационной безопасности и надежности работы не представляется возможным. Единственным вариантом является контроль разработчика, который также имеет ограниченные возможности. Это определяет необходимость производства необходимых для ИНС процессоров в России.

3. Практическая реализация доверенных систем управления с глубинным машинным обучением на базе имеющихся технологий для существенной части задач в режиме реального времени не может быть обеспечена, для другой части задач такая реализация связана с существенным падением производительности.

4. Повышение производительности доверенных систем управления может быть основано на совершенствовании архитектуры гибридных вычислительных машин, в том числе одновременном использовании процессоров разной архитектуры, оптимальных для решения соответствующих задач анализа и управления.

Обсуждение и заключения. В настоящей работе проведен анализ и рассмотрены вопросы значения и реализации машинного обучения в условиях цифровой трансформации промышленности. Затронутая в работе научная проблема заключается в недостаточном развитии технического уровня современных вычислительных машин для обеспечения высокой производительности алгоритмов на основе глубинного машинного обучения. Обращает на себя внимание проблема обеспечения информационной безопасности, что является одной из предпосылок развития отечественных процессоров для ИНС. На основе анализа требований к доверенным системам управления определены объективные сложности их реализации на базе существующих технологий и необходимость разработки новых подходов к архитектуре вычислительных машин.

Библиографический список

1. Using neuro-accelerators on FPGAs in collaborative robotics tasks / A. Zelensky, E. Semenishchev, A. Alepko [et al.] // SPIE Optical Instrument Science, Technology, and Applications II. — 2021. — Vol. 11876. – Art. 118760O. — P. 5. https://doi.org/10.1117/12.2600582

2. Zelenskii, A. A. Control of Collaborative Robot Systems and Flexible Production Cells on the Basis of Deep Learning / A. A. Zelenskii, M. M. Pismenskova, V. V. Voronin // Russian Engineering Research. — 2019. — Vol. 39. — P. 1065–1068. https://doi.org/10.3103/S1068798X19120256

3. Automated visual inspection of fabric image using deep learning approach for defect detection / V. V. Voronin, R. A. Sizyakin, M. Zhdanova [et al.] // Automated Visual Inspection and Machine Vision IV. — 2021. — Vol. 11787. — P. 117870. https://doi.org/10.1117/12.2592872

4. NeuFlow: Dataflow Vision Processing System-on-a-Chip / Phi-Hung Pham, D. Jelaca, C. Farabet [et al.] // In: Proc. IEEE 55th International Midwest Symposium on Circuits and Systems (MWSCAS). — 2012. — P. 1044–1047. https://doi.org/10.1109/MWSCAS.2012.6292202

5. Шуремов, Е. Л. Стоит ли увлекаться Большими Данными? / Е. Л. Шуремов // Учет. Анализ. Аудит. — 2020. — Т. 7, № 2. — С.17–29. https://doi.org/10.26794/2408-9303-2020-7-2-17-29

6. Towards artificial general intelligence with hybrid Tianjic chip architecture / Jing Pei, Lei Deng, Sen Song [et al.] // Nature. — 2019. — Vol. 572. — P. 106–111. https://doi.org/10.1038/s41586-019-1424-8

7. Модха, Д. TrueNorth: от нуля к 64 миллионам нейронов / Д. Модха // Открытые системы. СУБД. — 2019. — № 3. — С. 8.

8. TrueNorth: design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip / F. Akopyan, J. Sawada, A. Cassidy [et al.] // IEEE transactions on computer-aided design of integrated circuits and systems. — 2015. — Vol. 34. — P. 1537–1557. https://doi.org/10.1109/TCAD.2015.2474396

9. Loihi: A neuromorphic manycore processor with on-chip learning / Mike Davies, Narayan Srinivasa, Tsung-Han Lin [et al.] // IEEE Micro. — 2018. — Vol. 38. — P. 82–99. https://doi.org/10.1109/MM.2018.112130359

10. Efficient synapse memory structure for reconfigurable digital neuromorphic hardware / J. Kim, J. Koo, T. Kim, J. J. Kim // Frontiers in neuroscience. — 2018. — Vol. 12. — P. 829. https://doi.org/10.3389/fnins.2018.00829

11. Федоров, А. Квантовые вычисления: от науки к приложениям / А. Федоров // Открытые системы. СУБД. — 2019. — № 3. — С. 14.

12. What Happens When ‘If’ Turns to ‘When’ in Quantum Computing? / J. F. Bobier, M. Langione, E. Tao [et al.] // BCG Digital Transformation. — 2021. — P. 20.

13. Сабанов, А. Г. Доверенные системы как средство противодействия киберугрозам / А. Г. Сабанов // Защита информации. Инсайд. — 2015. — № 3 (63). — С. 17–21.

14. Каляев, И. А. Доверенные системы управления / И. А. Каляев, Э. В. Мельник // Мехатроника, автоматизация, управление. — 2021. — Т. 22, № 5. — С. 227–236. https://doi.org/10.17587/mau.22.227-236

Поступила в редакцию 27.12.2021
Поступила после рецензирования 17.01.2022
Принята к публикации 18.01.2022

Об авторах:

Зеленский Александр Александрович, директор Института цифровых интеллектуальных систем Московского государственного технологического университета «СТАНКИН» (127055, РФ, г. Москва, пер. Вадковский, 3 а), кандидат технических наук, доцент, Scopus, Researcher, ORCID, Zelenskyaa@gmail.com

Абдуллин Тагир Хабибович, преподаватель кафедры «Промышленная электроника и интеллектуальные цифровые системы» Московского государственного технологического университета «СТАНКИН» (127055, РФ, г. Москва, пер. Вадковский, 3 а), ведущий инженер, Scopus, ORCID, everestultimate@yandex.ru

Жданова Марина Михайловна, младший научный сотрудник ФГБОУ ВО «Московский государственный технологический университет «СТАНКИН» (127055, РФ, г. Москва, пер. Вадковский переулок, 3 а), Scopus, Researcher, ORCID, mpismenskova@mail.ru

Воронин Вячеслав Владимирович, заместитель директора Центра когнитивных технологий и машинного зрения Московского государственного технологического университета «СТАНКИН» (127055, РФ, г. Москва, пер. Вадковский, 3 а), кандидат технических наук, доцент, Scopus, Researcher, ORCID, voronin_sl@mail.ru

Грибков Андрей Армович, директор Аналитического центра Московского государственного технологического университета «СТАНКИН» (127055, РФ, г. Москва, пер. Вадковский, 3а), доктор технических наук, профессор, Scopus, Researcher, ORCID, andarmo@yandex.ru

Заявленный вклад соавторов:

А. А. Зеленский — формирование основной концепции, цели и задачи исследования, подготовка текста, формирование выводов; Т. Х. Абдуллин, М. М. Жданова — проведение исследования, анализ существующих подходов; В. В. Воронин, А. А. Грибков — анализ результатов исследований, доработка текста, корректировка выводов.

Все авторы прочитали и одобрили окончательный вариант рукописи.